图书情报工作
第64卷 第21期 2020年11月

基于二模复杂网络的隐性知识发现方法研究*
——以潜在药物靶点挖掘为例

■ 李东巧[1] 陈芳[1] 韩涛[1] 杨艳萍[1] 王学昭[1] 王燕鹏[1] Cynthia Liu[2] Yingzhu Li[2]
[1] 中国科学院文献情报中心 北京 100190 [2] 美国化学文摘社 哥伦布市 43202

摘 要：[目的/意义]通过构建二模复杂网络模型，揭示隐藏在海量文献中的隐性知识。[方法/过程]通过 NetworkX 复杂网络工具包，依据任意两个节点的共现关系构建二模复杂网络模型；对网络模型中节点的共现关系进行加权，计算网络的拓扑信息并进行 AP 聚类，提取节点间的直接关系；采用 AUC 方法对 AA、JC、加权改进的 wAA 和 wJC 等 4 种链路预测算法进行评价，遴选出最合适的预测算法，并对复杂网络的隐性关系进行预测分析。[结果/结论]以潜在药物靶点挖掘为例进行的实证研究结果表明，wAA 链路预测算法为最优的链路预测算法；二模复杂网络模型、指标和方法体系在美国化学文摘社数据库中的药物靶点挖掘中具有一定的有效性。下一步计划在其他数据库中或其他研究领域中进行尝试，以进一步验证该模型的通用性和有效性。

关键词：隐性知识 链路预测 复杂网络 药物靶点 疾病
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.21.015

1 引言

当今社会处于知识爆炸的大数据时代，大量的资料已经超出了人类对知识的吸收能力。随着各学科研究的不断交叉和融合，使得特定领域和主题的研究逐渐受到科研人员的关注和重视。知识从表现形式上有显性知识和隐性知识之分。相对于显性知识来说，隐性知识由于不易模仿和复制等特点[1]，成为研究人员不断创新的关键要素。如何挖掘隐藏在海量文献中的隐性知识是研究人员今后具备核心竞争优势的根本，也是当前大数据时代必须面临的机遇和挑战。

早期对隐性知识的挖掘主要基于"非相关文献的知识发现模型"，即 ABC 模型理论。ABC 模型是美国芝加哥大学的情报学教授 D. R. Swanson 于 1987 年提出的知识发现方法[2]，其基本思想是对两组非相关的文献集 A 和 C，如果一组文献表明 A 可以导致 B 发生，而另一组文献表明 B 可以导致 C 发生，那么通过逻辑递推关系，可推知 A 和 C 存在着一定逻辑联系。大量的文献集聚使研究内容彼此之间的关系呈现一种高度复杂性的网络，研究人员也可以通过知识网络对相关的隐性知识进行挖掘。

链路预测作为复杂网络数据挖掘领域的研究方向之一，主要利用现有的网络信息，预测已存在但尚未被发现的关系，或目前不存在但应该存在或者未来很可能存在的关系[2-4]。目前链路预测的研究方法包括基于邻居节点的 Admaic-Adar(AA)[5]、Jaccard(JC)[6]等，基于路径的 Katz[7]、FriendLink[8]等，以及基于随机游走的 Random walk with restart(RWR)[9]、Local random walk(LRW)[10]等。

链路预测在生物医学领域的挖掘已经取得了一系列的成果，主要集中在对非结构化的电子病历库、UniProt 等实验数据库、Web of Science 数据库、PubMed 数据库等中分析挖掘疾病与疾病[11-13]、基因与疾病[14]、基因与蛋白[15]、蛋白与蛋白[16-17]等的相互关联。但是

* 本文系中国科学院文献情报中心青年人才领域前沿项目"基于二模复杂网络的隐性知识发现方法研究-以潜在药物为例"(项目编号：G180181001)研究成果之一。

作者简介：李东巧(ORCID:0000-0002-7447-2436)，副研究员，博士，E-mail：lidq@mail.las.ac.cn；陈芳(ORCID:0000-0003-2517-5299)，副研究员，硕士；韩涛(ORCID:0000-0001-5955-7813)，研究员，博士；杨艳萍(ORCID:0000-0003-0428-4939)，副研究员，博士；王燕鹏(ORCID:0000-0002-2583-9895)，助理研究员，硕士；王学昭(ORCID:0000-0001-8496-3354)，副研究员，博士；Cynthia Liu(0000-0003-3858-1501)，科学信息经理，博士；Yingzhu Li(0000-0002-4946-7272)，信息科学家，博士。

收稿日期：2020-02-22 修回日期：2020-07-09 本文起止页码：120-129 本文责任编辑：杜杏叶

这些研究在推断关联关系中仍然存在一些局限性。首先，数据来源较为单一，此前的研究主要来源于病例库或专利库或论文库，但未将相关数据库的信息进行整合分析，预测信息会存在遗漏的可能；其次，关联节点有限，此前采用的研究方法未涉及在两个节点组成的网络中同类节点的关系。

美国化学文摘社数据库是全球最大的化学和相关学科信息的集成者，不仅包含论文数据和专利数据，还对收录的基因、蛋白质、药物等物质数据进行了标引。对多种来源的数据进行分析可以更完整地揭示整个领域的隐性知识。因此，本文采用二模复杂网络链路预测的方法，对美国化学文摘社中的论文、专利和物质等数据进行分析，深入挖掘其中的隐性知识，并以潜在药物靶点挖掘为实证研究，一定程度为节省新药研发时间和挖掘药物更多潜在的适应症提供一定参考。

2 研究思路与方法

本文从“构建二模复杂网络模型→提取复杂网络直接关系→遴选最优链路预测方法→预测复杂网络隐性关系”等4个方面出发，预测路线具体如图1所示：

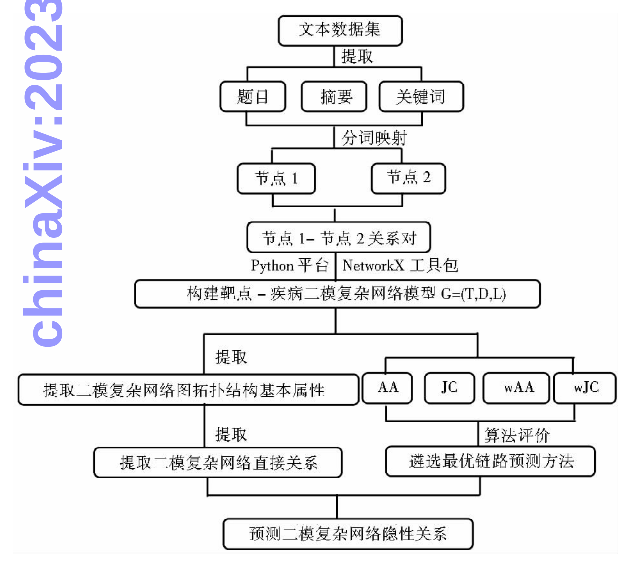


图1　二模复杂网络预测路线

2.1　构建二模复杂网络模型

在复杂网络中，节点代表真实世界中不同的个体，边代表个体间的关系，当两个不同节点之间具有某种特定关系时则连一条边，反之则不连边，其中，两个有边连接的节点在网络中称为邻接节点[18]。

与规则网络和随机网络不同，复杂网络呈现出高度的复杂性，主要体现在三个方面：①结构复杂性 复杂网络的连接结构非常复杂，可能随时发生变化；②节点复杂性 复杂网络的节点可能存在多种不同类型的节点；③各种复杂性因素的相互影响 对复杂网络来说，各种各样的因素可能都会产生不同的影响和作用，且网络与网络直接可能也存在某种联系[19]。

二模复杂网络是复杂网络的一种表示模式，由两种类型的节点构成。二模复杂网络模型可以用公式G＝(T, D, L)表示。其中，T和D分别代表任意两个关联的节点，L代表任意两个节点的关联关系。

本文采用的二模复杂网络与一般的二分网络有所不同。在二分网络中，同一类型的节点之间没有边相连，不同类型的节点才有边相连。在本研究中，L代表的关系为基于文本共现提取的相关性，同一类型的节点(如D1与D2，T1与T2)之间必然存在文本的共现相关性，且该相关性对于节点之间的隐性关联来说具有重要作用。因此，本文的研究依然保留同类节点的关系，只是引入一个参数达到对不同类型节点的连边取不同权重的目的。

2.2　提取二模复杂网络直接关系

复杂网络拓扑结构的基本属性与预测方法的性能密不可分，以下对其进行简单介绍[20]。网络效率指的是全部节点间的距离倒数和的平均值；节点度指的是网络中一个节点直接连接的节点个数；平均聚集系数指的是网络中包含任意一个节点的三角形结构比例的平均值；同配系数指的网络度和度的相关性，用来衡量节点之间连接的倾向性；平均度指的是对网络中所有节点的度值求平均值。

复杂网络中直接关系的提取主要通过利用SimRank相似度计算网络图的方式计算，并根据相关结果进行AP聚类，以提取节点间已知关系特征。其中，SimRank相似度指的是如果两个节点所连接的节点相似，那么这两个节点就相似，其基于网络图的拓扑结构，利用递归的定义方式可以捕捉到网络图结构的整体信息[21]。

与传统的文本相似度相比，SimRank相似度的计算完全基于网络图的拓扑结构，其递归的定义方式能使SimRank相似度的值捕捉到图结构的整体信息。与Google的PageRank算法只能衡量每个结点的重要性相比，SimRank相似度能比较任意两个结点间的相似度问题，因此，SimRank在计算该网络的相似矩阵方法具有一定的优势。

其中，SimRank的计算公式为 $S = C \cdot (W^T \cdot S \cdot W) + (1-C) \cdot I$。该公式中，S为相似度矩阵，W为邻接矩阵，$C$为衰减因子，I为单位矩阵。在本文实验

中,W 为加权的邻接矩阵,迭代次数为 20。

2.3 直接关系的 AP 聚类

AP 聚类指的是按照一定的规则在节点之间传递信息,在多次迭代过程中出现聚类中心,进而实现数据点的自动聚类,具有聚类速度快、对输入相似度矩阵的三角不等式和对称性没有要求、且可以适用于多种场合等优点[22-24]。AP 聚类的优势是可以不用人为设置初始的类中心,主要根据相似矩阵本身的特征逐渐达到聚类收敛的效果。根据 SimRank 计算出相似矩阵,构建节点的向量,计算公式为 Node_veci = [S_{i1}, S_{i2}, ⋯ S_{ij}, ⋯, S_{in}], i, j ∈ N。

AP 聚类算法源于 sklearn 工具包,其主要参数包括 affinity = 'euclidean', convergence_iter = 15, copy = True, damping = 0.5, max_iter = 200, preference = None, verbose = False。AP 聚类的过程自动产生类簇的数量,一开始其类簇数量超过 400 个,在可视化之前,进一步将类簇中心进行迭代聚类,获得类簇数量为 28 个。

2.4 遴选最优链路预测方法

链路预测是复杂网络研究的一个重要方向,主要通过网络中已知节点的节点度、节点对之间的路径、网络的平均最短距离、网络的簇系数等拓扑结构信息对网络进行相似性度量,预测复杂网络中尚未产生连接的 2 个节点之间产生连接的可能性[25-27]。本文基于邻居节点的链路预测方法,通过引入加权参数 α 对 AA、JC 等算法进行加权处理,其中,二模网络中的同类型节点和不同类型节点也均进行了加权处理。

其中,4 种算法的具体公式如下:

(1) $AAuv = \sum_{\omega \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(\omega)|}$

(2) $wAA_{uv} = \sum_{\omega \in \Gamma(u) \cap \Gamma(v)} \frac{w(u,\omega)^{\alpha} + w(v,\omega)^{\alpha}}{\log(1 + w|\Gamma(\omega)|)}$

(3) $JCuv = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$

(4) $wJC_{uv} = \frac{w|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} = \frac{\sum_{\omega \in \Gamma(u) \cap \Gamma(v)} w(u,\omega)^{\alpha} + w(v,\omega)^{\alpha}}{w|\Gamma(u)| + w|\Gamma(v)|}$

其中,函数 $\Gamma(u)$ 表示 u 的邻居节点,$w|\Gamma(u)| = \sum_{\omega \in \Gamma(u)} w(w,x)^{\alpha}$。α 为加权参数,当 $(\omega,x) \in (T,D)$,α = 1;当 $(\omega,x) \in (T,T)$ 或者 (D,D),α = 0。

AA 算法是指将不同的权重分配给该公共邻居集合中的不同节点,每个节点的权重等于该节点的度的对数分之一;wAA 算法是指引入加权参数 α 之后的 AA 算法,参数 α 的取值使得同类型节点的权重下降,不同类型节点的权重保持与原始共现值呈正比;JC 算法是指两个节点的共同邻居节点占两个节点邻居节点总和的比例来表示节点的相似性;wJC 算法是指引入加权参数 α 之后的 JC 算法,参数 α 作用与上述在 wAA 算法时一样。

2.5 链路预测方法的评价

本文采用 AUC (area under the receive operation characteristic Curve) 方法对 AA、JC、加权改进的 wAA 和 wJC 等 4 种链路预测算法进行评价。AUC 是指 ROC 曲线(receiver operating characteristic curve)下的面积,是衡量链路预测算法精确度最常用的一种标准。AUC 从整体上衡量算法的精确度,采用十倍交叉验证的方法,从有边的关系对中随机提取 10% 后任意切割为十份,包括九份训练集和一份为测试集,随后对测试集进行预测。将十次预测后的结果求平均得到 AUC,该指标通过公式 $AUC = (n' + 0.5n'')/n$ 计算。其中,n 为比较次数,n' 为测试集边的预测值大于不存在边的次数,n'' 为测试集边预测值等于不存在边的次数[28]。

3 实验结果分析

近年来,随着科研人员对疾病机理的深入了解和技术手段的不断进步,靶向药物治疗发挥着越来越重要的作用,因此靶点研究也成为了新药开发的重要研究方向。本文通过构建靶点-疾病二模复杂网络,根据已知靶点与疾病的关系,预测治疗疾病的尚未被发现的其他有效靶点,一定程度上将为提高新药研发进程、节省研发开支和降低研发风险提供一定的参考。

3.1 数据来源和加工

本文以美国化学文摘社数据库为文献来源,从中提取出与抗体药物相关的 514 539 篇论文和专利,以及其包含的抗体物质,数据获取日期截至 2018 年上半年。对相关文献数据进行深度标引,将其中涉及的疾病、靶点、物质,以及其他的标签进行人工清洗和合并,形成 1 015 个抗体靶点、3 867 种疾病标签库。其中,肿瘤类疾病是疾病标签库中最大的一个分支,数量为 2 137种(部分肿瘤节点与其他疾病种类有重叠,总数中已经去重)。

3.2 构建二模复杂网络模型

本文采用靶点和疾病两个关联词为节点,分别命名为 T_i 和 D_j,将二者在文献中的共现关系命名为 L_k。基于 Python 语言,利用 NetworkX 工具包构建二模复杂网络模型,并用公式 $G = (T_i, D_j, L_k)$ 表示。其中,根

据3.1的数据加工结果，T为1 015，D为3 867，L为911 479。

利用NetworkX工具包，对靶点-疾病网络拓扑结构进行分析。依据任意靶点T_i和疾病D_j两个节点的共现词频对节点间的关系进行加权，计算网络节点的节点数、边、效率、平均聚集系数、加权聚集系数、同配系数和平均度等，如表1所示：

表1　靶点-疾病二模复杂网络拓扑结构基本属性

节点数($T+D$)	边	效率	平均聚集系数	加权聚集系数	同配系数	平均度
4 882	911 479	0.387 9	0.666 1	$6.158\ 7\times10^{-4}$	-0.293 4	373.4

3.3　提取二模复杂网络直接关系

利用SimRank相似度计算网络图的拓扑信息，采用AP聚类对相似矩阵进行聚类，形成包含28个类簇的靶点-疾病二模复杂网络直接关系聚类图。其中，每种颜色代表一个类簇，如图2所示：

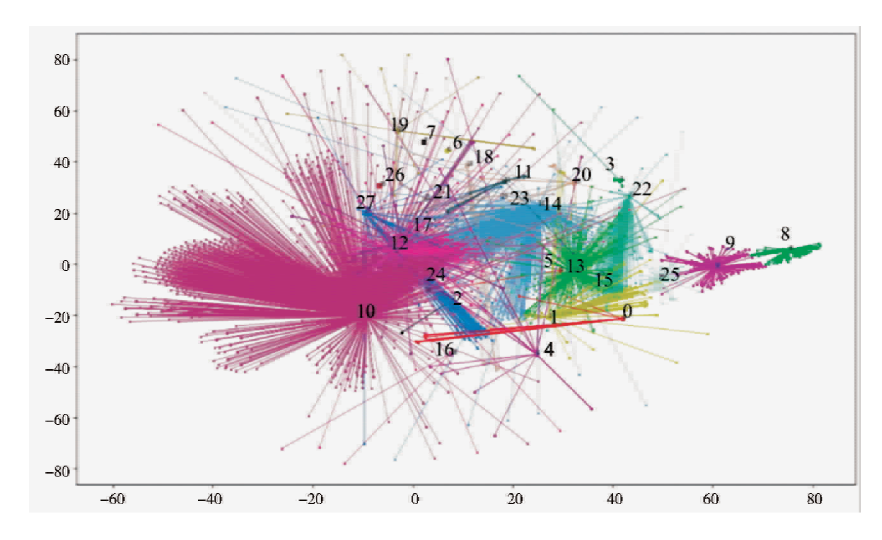


图2　靶点-疾病二模复杂网络直接关系聚类图

类簇节点数量大于500的巨型类簇共有两个，分别是类簇10(2 741个节点)和类簇14(563个节点)(见表2)。类簇10是靶点-疾病的关系对数量最多的类簇，共包含672个靶点节点和2 069个疾病节点。在类簇10的靶点节点中，与疾病存在关联的数量最多的三个靶点分别是epidermal growth factor receptor、carcinoembryonic antigen和Notch ligand DLL4；在类簇10的疾病节点中，与靶点存在关联的数量最多的三种疾病分别是肌腱炎(Tendinitis)、血细胞减少(Cytopenia)和原发性硬化性胆管炎(Primary sclerosing cholangitis)。类簇14是靶点-疾病的关系对数量位列第二位的类簇，共包含77个靶点节点和486个疾病节点。其中，在类簇14的靶点节点中，与疾病存在关联的数量最多的三个靶点分别是CD80 antigen、Ganglioside GD3和Tumor-associated glycoprotein 72；在类簇14的疾病节点中，与靶点存在关联的数量最多的三种疾病分别是自闭症(Autism)、先天性心脏病(Congenital heart disease)和严重联合免疫缺陷病(Severe combined immunodeficiency)。

3.4　遴选最优链路预测方法

表3对各预测算法所计算的关系对中，预测值大于0的关系对以及靶点和疾病的数量进行统计。其中，疾病代表所有疾病的全局网络，肿瘤代表肿瘤类疾病的子网络。肿瘤网络是在全局网络预测后，提取出来与肿瘤相关的局部网络。首先，wAA与AA，wJC与JC在预测值的数量上分别相同，表明加权后的算法不会遗漏靶点和疾病的统计；其次，AA(wAA)与JC(wJC)之间是有差异的，在疾病网络中的差异较小，在肿瘤网络中的差异比较明显。在肿瘤网络中，AA(wAA)在预测的关系对数量高于JC(wJC)，但在疾病数量中低于后者。

从表4可以看出，采用AUC方法对AA、JC、wAA和wJC等4种链路预测算法进行评价，wAA算法的AUC值最高(0.971 4)，为最优的链路预测算法，因此本文的实证分析是基于wAA算法的结果进行分析。

表 2 TOP 28 靶点 – 疾病的 AP 聚类分布

类簇标号	节点总数	靶点节点数量	疾病节点数量	TOP 3 靶点数量节点	TOP 3 疾病数量节点
0	14	4	10	CD3 antigen; Blood-coagulation factor IX; Blood-coagulation factor X	Neuroendocrine system neoplasm; Blood coagulation disorders; Hemophilia A
1	75	25	50	Sialic acid-binding Ig-like lectin 2; CD4 antigen; FcεRI receptor	Macular degeneration; Epilepsy; Irritable bowel syndrome
2	6	6	0	Monocyte chemoattractant protein-1; Chemokine CXCL12; RANTES (chemokine)	–
3	9	1	8	Glucagon-like peptide I	Metabolic syndrome X; Hyperglycemia; Sleep disorders
4	24	12	12	Angiopoietin 2; Paliperidone; Heparin-binding EGF-like growth factor	Cerebral palsy; Arterial thrombosis; Decubitus ulcer
5	2	1	1	Gingipain R	Periodontal disease;
6	4	3	1	Viral envelope glycoprotein UL130; Viral envelope glycoprotein H; Viral envelope glycoprotein L	Vascular endothelium disease
7	2	0	2	–	Chronic hepatitis LT hepatitis B; Acute hepatitis B
8	66	5	61	CD20 antigen; Tyrosine kinase receptor HER2; Vascular endothelial growth factor	Rheumatoid arthritis; Mammary gland neoplasm; Autoimmune disease
9	120	3	117	Integrin αM; Sialic acid-binding Ig-like lectin 3; Cytotoxic T-lymphocyte-associated protein 4	Hypertension ; Parkinson disease; Myocardial infarction
10	2741	672	2069	epidermal growth factor receptor; carcinoembryonic antigen; Notch ligand DLL4	Tendinitis; Cytopenia; Primary sclerosing cholangitis
11	16	2	14	Ki-67 antigen; Transferrin receptor	Fertility disorders LT male; Familial adenomatous polyposis; Atrophic gastritis
12	487	63	424	Interleukin 1α; B7 homolog 3 protein; Sphingosine 1-phosphate	Kidney injury; Trypanosomiasis; Gallbladder disease
13	257	13	244	Hepatocyte growth factor; Epidermal growth factor receptor HER4; Interleukin 4	Hepatic fibrosis; Hypothyroidism; Nerve disease
14	563	77	486	CD80 antigen; Ganglioside GD3; Tumor-associated glycoprotein 72	Autism; Congenital heart disease; Severe combined immunodeficiency
15	1	0	1	–	Pemphigus foliaceus
16	4	3	1	T cell receptor αβ; Cathepsin K; L-Lactate dehydrogenase	Immunosuppression LT cellular
17	60	11	49	Advanced glycosylation end product receptor; Interleukin 33; Langerin	Charcot-Marie-Tooth disease; Hypertrophic cardiomyopathy; Aortic stenosis
18	29	15	14	Lymphocyte activation gene-3 protein; Integrin αVβ6; CD28 antigen	Acute B-cell leukemia; Influenza type A; Kidney ischemia
19	21	15	6	Fibroblast growth factor receptor 1; β-Klotho protein; B cell receptor	Thrombus; Wound infection; Atherothrombosis
20	18	7	11	Neural apoptosis-regulated convertase 1; MADCAM-1 protein; Zaire ebolavirus	Hypercholesterolemia; Nerve injury; Pancreatic adenocarcinoma
21	17	4	13	Blood serum albumin; Tumor necrosis factor receptor 1; Human albumin	Cerebrovascular disease; Polycystic ovary syndrome; Ovarian disease
22	132	3	129	β-Amyloid; Mucin 1; Integrin αV	Muscle disease; Cardiac arrhythmia; Arteriosclerosis
23	4	1	3	Insulin receptor	Hyperglycemia LT glucose intolerance; Hyperinsulinemia; Hyperphagia
24	8	1	7	Vascular endothelial growth factor B	Alport syndrome; Albuminuria; Osteodystrophy
25	104	7	97	Interleukin 2 receptor; Interleukin 6 receptor; Integrin α4	Vascular disease; Hepatitis C; Traumatic injury
26	2	0	2	–	Trisomy; Human trisomy 8 syndrome
27	94	65	29	Bone morphogenetic proteins, sclerostin; Prostaglandin E2; P-selectin	Erysipelas; Stevens-Johnson syndrome; Erythroblastosis fetalis

表 3 预测结果的节点和关系对数量

算法	AA (wAA)		JC(wJC)	
	疾病	肿瘤	疾病	肿瘤
预测的关系对(预测值 >0)	2 544 973	393 149	2 544 973	370 649
涉及靶点数	1 015	1 015	1 015	1 015
涉及疾病数	3 726	515	3 710	996

表 4 二模复杂网络的指标评价

指标	AA	wAA	JC	wJC
AUC	0.948 5	0.971 4	0.882 6	0.969 8

3.5　预测二模复杂网络隐性关系

本文预测的所有关系对超过200万条，在预测值越高关系对成立可能越大的基础上，利用NetworkX工具包，以wAA算法的结果为依据，筛选TOP100靶点－疾病的关系对，将靶点－疾病的复杂网络关系进行可视化展示（见图3）。其中，红色圆圈代表靶点，蓝色方框代表治疗的疾病，绿色实线代表靶点与疾病之间的直接关系，即已知某个靶点与治疗某种疾病存在关联关系，紫色虚线代表靶点与疾病之间的隐性关系，即某个靶点可能与治疗某种疾病存在关联关系，线条的粗细代表关系的强弱。

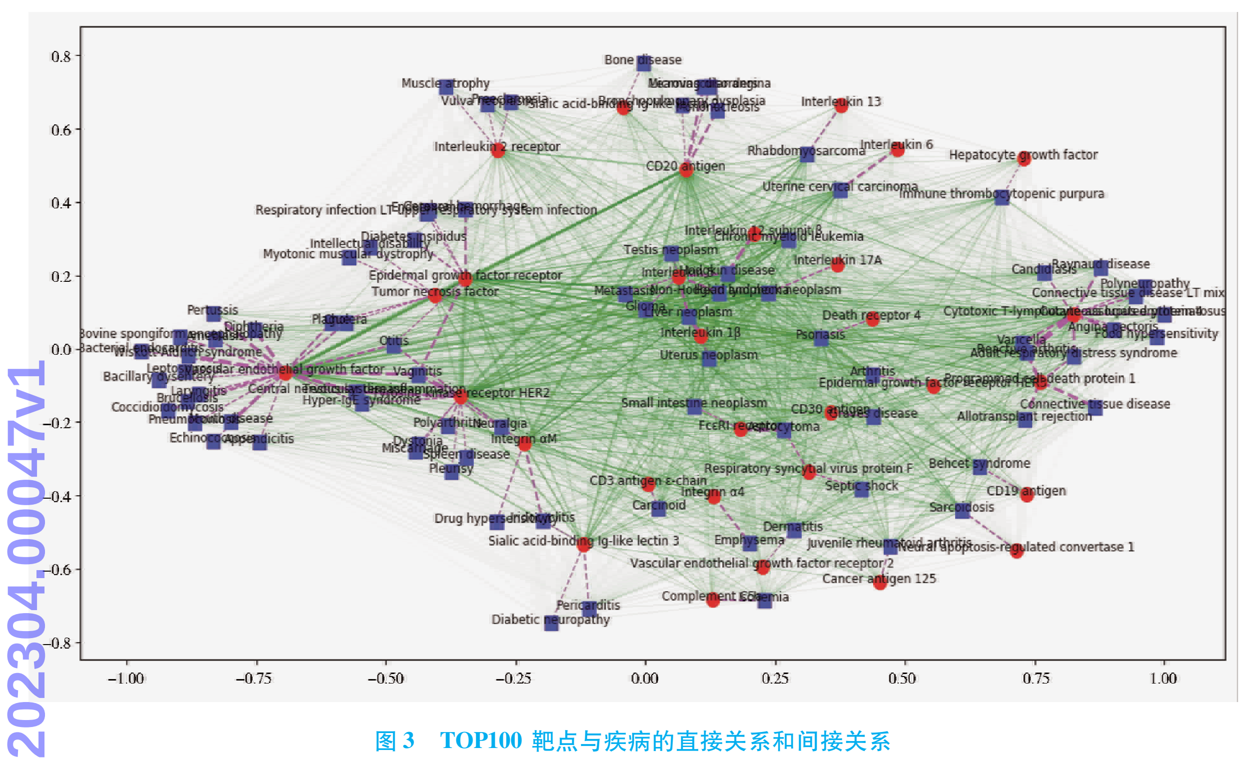


图3　TOP100靶点与疾病的直接关系和间接关系

由图中可以看出，在TOP100靶点－疾病关系中，直接关系最为密切的5个关系对分别是：表皮生长因子受体（Epidermal growth factor receptor）和CD20抗原（CD20 antigen）、表皮生长因子受体和血管内皮生长因子（Vascular endothelial growth factor）、表皮生长因子受体和Tyrosine kinase receptor HER2、CD20抗原和Tyrosine kinase receptor HER2、CD20抗原和Integrin αM等。

细胞毒性T淋巴细胞相关蛋白4（Cytotoxic T-lymphocyte-associated protein 4，CTLA 4）和程序性细胞死亡蛋白1（Programmed cell death protein 1，PD-1）能够增强特异性抗肿瘤的免疫反应，是免疫检查点疗法的两个重要靶点，且James P. Allison教授因发现了针对CTLA-4的免疫检查点疗法和Tasuku Honjo教授（发现PD-1免疫检查点疗法）一同获得2018年诺贝尔生理学或医学奖。从TOP20靶点－疾病的间接关系对中可以看出，以CTLA 4为靶点的药物治疗反应性关节炎（Reactive arthritis）的预测值最高，预测其还可以治疗心绞痛（Angina pectoris）、成人呼吸窘迫综合征（Adult respiratory distress syndrome）和雷诺综合征（Raynaud disease）等，相关预测值分别排名为第12位、17位和19位。以PD-1为靶点的药物治疗反应性关节炎（Reactive arthritis）的预测值排名第2位，预测其还可以治疗同种异体移植排斥反应（Allotransplant rejection）和结缔组织疾病（Connective tissue disease），相关预测值分别排名为第13位和第14位。

血管内皮生长因子（Vascular endothelial growth factor）也是抗肿瘤药物的重要靶点之一。从表5可以看出，以血管内皮生长因子为靶点的药物预测治疗疾病的种类最多，包括Wiskott-Aldrich氏症候群（Wiskott-AldrichSyndrome）、白喉（Diphtheria）、阴道炎（Vaginitis）、口腔疾病（Mouth disease）、百日咳（Pertussis）和中枢神经系统炎症（Central nervous system inflammation）等。

由于在靶点－疾病隐性关系预测中未能展示较多靶点－肿瘤的隐性关系，加上肿瘤在疾病标签库中所占比例较高，因此将靶点－肿瘤的隐性关系单列出来进行分析。TOP100靶点与肿瘤的直接关系和隐性关系见图4。从图4中可以看出，与肿瘤直接关系最为密切的前3个靶点分别是CD20抗原（CD20 antigen）、酪氨酸激酶受体HER2（Tyrosine kinase receptor HER2）和血管内皮生长因子（Vascular endothelial growth factor）。

表 5 TOP20 靶点－疾病隐性关系预测

序号	预测值	靶点	疾病
1	0. 252 176	Cytotoxic T-lymphocyte-associated protein 4	Reactive arthritis
2	0. 227 286	Programmed cell death protein 1	Reactive arthritis
3	0. 224 872	Tyrosine kinase receptor HER2	Vaginitis
4	0. 224 487	Vascular endothelial growth factor	Wiskott-Aldrich syndrome
5	0. 223 732	Interleukin 1β	Liver neoplasm
6	0. 223 272	Vascular endothelial growth factor	Diphtheria
7	0. 220 884	FcεRI receptor	Astrocytoma
8	0. 220 318	Vascular endothelial growth factor	Vaginitis
9	0. 220 090	Interleukin 12 subunit β	Hodgkin disease
10	0. 219 635	Vascular endothelial growth factor	Mouth disease
11	0. 217 389	Vascular endothelial growth factor	Pertussis
12	0. 216 271	Cytotoxic T-lymphocyte-associated protein 4	Angina pectoris
13	0. 215 814	Programmed cell death protein 1	Allotransplant rejection
14	0. 214 442	Programmed cell death protein 1	Connective tissue disease
15	0. 211 445	Tyrosine kinase receptor HER2	Central nervous system inflammation
16	0. 211 154	Vascular endothelial growth factor	Central nervous system inflammation
17	0. 209 907	Cytotoxic T-lymphocyte-associated protein 4	Adult respiratory distress syndrome
18	0. 208 939	CD20 antigen	Learning disorders
19	0. 208 375	Cytotoxic T-lymphocyte-associated protein 4	Raynaud disease
20	0. 206 707	Interleukin 6	Uterine cervical carcinoma

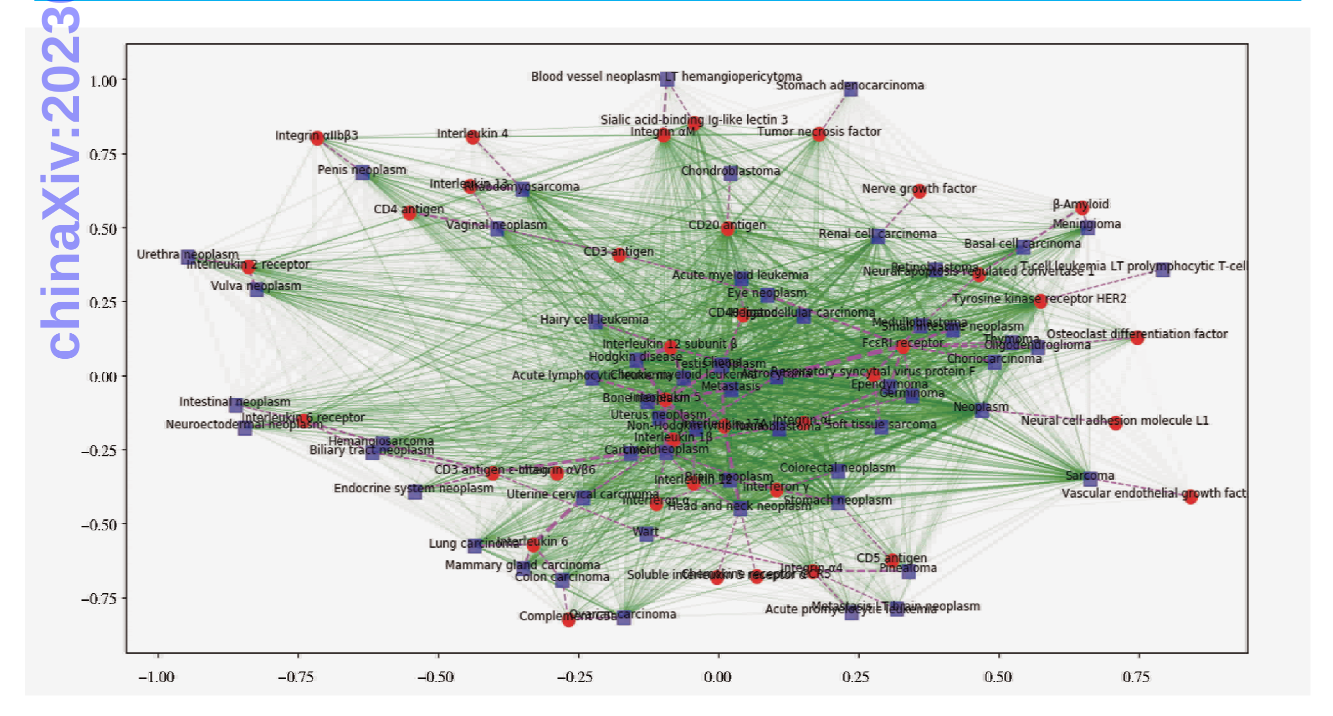


图 4 TOP100 抗体靶点与肿瘤的直接关系和间接关系

Interleukin 1β 是治疗自身免疫性疾病和阿尔兹海默等疾病的重要靶点。从 TOP20 靶点－肿瘤的间接关系对中可以看出(见表 6),以 Interleukin 1β 为靶点的药物治疗肝癌(Liver neoplasm)的预测值最高,预测其还可以治疗子宫肿瘤(Uterus neoplasm),相关预测值排名第 18 位。FcεRI 受体(FcεRI receptor)是治疗变态反应性疾病等重要靶点之一。从表 5 可以看出,以 FcεRI 受体为靶点的药物治疗星形细胞瘤(Astrocytoma)的预测值排名第 2 位,预测其还可以治疗小肠肿瘤(Small intestine neoplasm)、室管膜瘤(Ependymoma)

和眼部肿瘤(Eye neoplasm)等疾病，相关预测值分别排名第13位、第19位和第20位。

白介素5(Interleukin 5)是治疗支气管哮喘等疾病等重要靶点，且以白介素5为靶点的药物预测治疗疾病的种类最多，包括肝癌(Liver neoplasm)、非霍奇金淋巴瘤(Non-Hodgkin lymphoma)、霍奇金淋巴瘤(Hodgkin disease)、神经胶细胞瘤(Glioma)、转移瘤(Metastasis)和睾丸肿瘤(Testis neoplasm)等。

表6　TOP20靶点－肿瘤隐性关系预测

序号	预测值	靶点	肿瘤
1	0.223 732	Interleukin 1β	Liver neoplasm
2	0.220 884	FcεRI receptor	Astrocytoma
3	0.220 090	Interleukin 12 subunit β	Hodgkin disease
4	0.206 707	Interleukin 6	Uterine cervical carcinoma
5	0.203 838	Interleukin 5	Liver neoplasm
6	0.198 955	Interleukin 5	Non-Hodgkin lymphoma
7	0.198 914	Interleukin 5	Hodgkin disease
8	0.192 688	Interleukin 13	Rhabdomyosarcoma
9	0.190 067	Respiratory syncytial virus protein F	Astrocytoma
10	0.188 85	Interleukin 17A	Head and neck neoplasm
11	0.185 449	CD3 antigen ε-chain	Carcinoid
12	0.181 121	Interleukin 12 subunit β	Chronic myeloid leukemia
13	0.180 465	FcεRI receptor	Small intestine neoplasm
14	0.179 437	Interleukin 5	Glioma
15	0.178 568	Interleukin 2 receptor	Vulva neoplasm
16	0.178 427	Interleukin 5	Metastasis
17	0.178 146	Interleukin 5	Testis neoplasm
18	0.177 031	Interleukin 1β	Uterus neoplasm
19	0.174 786	FcεRI receptor	Ependymoma
20	0.171 939	FcεRI receptor	Eye neoplasm

为了进一步验证预测结果，本文以“CTLA 4为靶点的药物预测治疗反应性关节炎、心绞痛、成人呼吸窘迫综合征和雷诺综合征”为例，以Pubmed论文数据库和Incopat专利数据库为验证数据库，分别在题目和摘要中进行信息检索，时间范围限定在入库时间至2018年6月。由表7可以看出，通过相关信息检索，2018年6月之前，在PubMed论文数据库和IncoPat专利数据库中存在关于CTLA4相关分子或蛋白在反应性关节炎、心绞痛、成人呼吸窘迫综合征和雷诺综合征中的机理和临床等研究，但是这些论文或专利未提及选用CTLA4作为抗体靶点进行相关疾病治疗。验证结果在一定程度上说明通过wAA链路预测算法能够识别现有公开论文或专利中未报道的隐性知识关联，为从事相关领域研究的科研人员提供参考。

表7　靶点－疾病隐性预测关系验证

靶点	疾病	预测关系	PubMed 数据库检索	IncoPat 数据库检索	信息检索结果分析
CTLA 4	Reactive arthritis	预测以CTLA 4为靶点的药物治疗反应性关节炎	0	2	共检索到对比文献2篇，主要关于不同结合蛋白在临床中治疗反应性关节炎等多种疾病的研究，未提及选用CTLA4作为抗体靶点治疗反应性关节炎的研究
	Angina pectoris	预测以CTLA 4为靶点的药物治疗心绞痛	1	2	共检索到对比文献3篇，主要关于CTLA4相关分子或蛋白治疗心绞痛的研究，未提及选用CTLA4作为抗体靶点治疗心绞痛的研究
	Adult respiratory distress syndrome	预测以CTLA 4为靶点的药物治疗成人呼吸窘迫综合征	2	0	共检索到对比文献2篇，主要关于呼吸窘迫综合征的机理研究，以及CTLA4相关药物治疗1名癌症患者的研究(该患者也患有严重急性呼吸窘迫综合征)，但是未提及选用CTLA4作为抗体靶点专门治疗成人呼吸窘迫综合征的研究
	Raynaud disease	预测以CTLA 4为靶点的药物治疗雷诺综合征	0	1	共检索到对比文献1篇，主要关于不同结合蛋白在临床中治疗雷诺综合征等多种疾病的研究，但是未提及选用CTLA4作为抗体靶点治疗雷诺综合征的研究

4 讨论

本文基于 Python 平台，利用 NetworkX 软件包构建二模复杂网络模型，选用改进的 wAA 链路预测算法对模型进行分析，并以药物靶点预测为实证分析，能够有效揭示靶点 - 疾病二模复杂网络中潜在的药物治疗靶点，研究结果为节省新药研发时间和挖掘药物更多潜在的适应症提供一定参考。

但是本文的研究工作仍有待改进：①在进行链路预测时，本文从基于邻居节点的链路预测算法进行研究，未涉及对基于路径和基于随机游走的算法，下一步将采用其他链路预测算法进行尝试。②在进行实证研究时，由于疾病种类具有一定的包含关系，因此在预测时是否需要将同一类型的疾病进行合并。如果同一类型的疾病进行合并，在预测时可能会将未知关系作为已知关系处理，这将漏掉一些研究人员非常关注且很有价值的细节领域的隐性关联；如果同一类型的疾病不进行合并，在预测时可能会将一部分已知关系作为未知关系处理，这将给未知关系带来了一些“噪音”，这些“噪音”的处理需要非常专业的研究人员进行人工判读，且剔除噪音后期的工作量太大，如何有效的改进还需要进一步讨论。③本文将二模复杂网络模型在潜在药物靶点挖掘上进行了实证研究，下一步计划在其他研究领域或其他数据库中进行比较，以进一步验证该模型的通用性和有效性。

致谢：本文在数据分析方面还得到了美国化学文摘社 Yi Deng、马清扬、余敏等人员的帮助和指导，在此表示感谢。

参考文献：

[1] 周青玲. 用户隐性知识的挖掘流程及实现技术[J]. 中国科技信息, 2015(11): 61 - 62.

[2] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社. 2013.

[3] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651 - 661.

[4] 姚亚兵. 基于复杂网络拓扑结构的链路预测方法研究[D]. 兰州:兰州大学, 2017.

[5] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social networks, 2003, 25(3):211 - 230.

[6] JACCARD, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. Bulletin de la societe vaudoise des sciences naturelles, 1901, 37:547 - 579.

[7] KATZ L A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1):39 - 43.

[8] PAPADIMITRIOU A, SYMEONIDIS P, MANOLOPOULOS Y. Fast and accurate link prediction in social networking systems[J]. Journal of systems and software, 2012, 85(9):2119 - 2132.

[9] BRIN S, PAGE L. Reprint of: the anatomy of a large-scale hypertextual Web search engine[J]. Computer networks, 2012, 56(18):3825 - 3833.

[10] LIU W, LUE L. Link prediction based on local random walk[J]. Epl, 2010, 89(5):58007.

[11] 余黄樱子, 董庆兴, 张斌. 基于网络表示学习的疾病知识关联挖掘与预测方法研究[J]. 情报理论与实践, 2019, 42(12): 156 - 162.

[12] 李星. 基于复杂网络的症状基因预测方法研究[D]. 北京: 北京交通大学, 2014.

[13] BUKET K, MUSTAFA P. Age-series based link prediction in evolving disease networks[J]. Computers in biology and medicine, 2015, 63:1 - 10.

[14] 丁亮. 基于异质性网络链路预测算法的非编码 RNA - 疾病相关性预测研究[D]. 安徽:中国科学技术大学, 2018.

[15] HU H, ZHU C Y, AI H X. LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction[J]. Molecular biosystems, 2017, 13(9): 1781 - 1787.

[16] 吴金华. 基于数据挖掘的阿尔兹海默症蛋白质网络研究[D]. 沈阳:辽宁大学, 2018.

[17] CRICHTON G, GUO Y F, PYYSALO S. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches[J]. BMC bioinformatics, 2018, 19:176.

[18] 周涛, 柏文洁, 汪秉宏, 等. 复杂网络研究概述[J]. 物理, 2005, 34(1):31 - 36.

[19] 李星. 基于复杂网络的症状基因预测方法研究[D]. 北京: 北京交通大学, 2014.

[20] 李兰茜. 基于复杂网络结构的链路预测技术研究[D]. 北京: 北京邮电大学, 2019.

[21] 张斌, 李亚婷. 学科合作网络链路预测结果的排序鲁棒性[J]. 信息资源管理学报, 2018, 8(4):89 - 97.

[22] 葛军. 一种重叠社区发现算法及其在 MapReduce 上的实现[D]. 西安:电子科技大学, 2013.

[23] FREY B J, DELBERT D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814):972 - 976.

[24] 王林, 董小江. 社团挖掘的并行化 AP 聚类方法[J]. 微型机与应用, 2017, 36(12):16 - 18.

[25] LU L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica a: statistical mechanics and its applications, 2011, 390(6):1150 - 1170.

[26] 杨晓翠, 宋甲秀, 张曦煌. 基于网络表示学习的链路预测算法[J]. 计算科学与探索, 2019, 13(5): 812 - 821.

[27] 杨育捷. 复杂网络下基于拓扑相似性的链路预测研究[D]. 北京: 北京邮电大学, 2019.

[28] 陈嘉颖, 于炯, 杨兴耀, 等. 基于复杂网络节点重要性的链路预测算法[J]. 计算机应用, 2016, 36(12):3251 - 3255,3268.

李东巧，陈芳，韩涛，等.基于二模复杂网络的隐性知识发现方法研究——以潜在药物靶点挖掘为例[J].图书情报工作，2020,64(21):120－129.

作者贡献说明：

李东巧：论文思路构建，整理数据并进行统计分析，撰写论文；

陈芳：参与论文研究方案设计，代码实现，整理研究方法，参与整理数据；

韩涛：指导论文研究方案；

杨艳萍：指导论文研究方案；

王学昭：指导论文研究方案；

王燕鹏：参与研究方法设计；

Cynthia Liu：参与数据加工；

Yingzhu Li：参与数据加工。

Research on the Tacit Knowledge Discovery Based on Two-mode Complex Network ——Take mining Potential Drug Targets as an Example

Li Dongqiao[1] Chen Fang[1] Han Tao[1] Yang Yanping[1] Wang Xuezhao[1]

Wang Yanpeng[1] Cynthia Liu[2] Yingzhu Li[2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Chemical Abstracts Service, Columbus, OH 43202, USA

Abstract: [**Purpose/significance**] This paper aims to extract the tacit knowledge from the massive literatures by constructing a two-mode complex network model. [**Method/process**] Through the NetworkX complex network toolkit, a two-mode complex network model was constructed based on the co-occurrence relationship of any two nodes. The direct relationship between nodes and nodes was extracted by weighting the co-occurrence relationship of nodes in the network model, calculating the topology information of the network and AP clustering. The most appropriate prediction algorithm was selected by using AUC method to evaluate the four link prediction algorithms, such as AA, JC, wAA and wJC. The tacit knowledge was predicted by the most appropriate prediction algorithm from the complex networks. [**Result/conclusion**] The results showed that the wAA link prediction algorithm was the optimal link prediction algorithm. The two mode complex network model, indicators and method system were effective in drug target mining in the Chemical Abstracts Service database. The next step is to try in other databases or other research fields to further verify the generality and effectiveness of the model.

Keywords: tacit knowledge link prediction complex network drug target diseases